Hy Murveit, SRI International
(415) 859-5447, hy@speech.sri.com,
High Performance Speech Recognition Using Consistency Modeling,
N00014-92-C-0154, 1 Oct 92 - 30 Sep 93

1. PRODUCTIVITY MEASURES

Refereed papers submitted but not yet published	2
Refereed papers published	2
Invited presentations	8

Hy Murveit, SRI International
(415) 859-5447, hy@speech.sri.com,
High Performance Speech Recognition Using Consistency Modeling,
N00014-92-C-0154, 1 Oct 92 - 30 Sep 93

2. DETAILED SUMMARY OF TECHNICAL PROGRESS

SRI's consistency modeling project began in August 1992. The goal of the project is to develop *consistency modeling* technology. Consistency modeling aims to reduce the number of improper independence assumptions used in traditional speech recognition algorithms, so that the resulting speech recognition hypotheses are more self-consistent and, therefore, more accurate. Consistency is achieved by conditioning HMM output distributions on state and observations histories, $P(x/s,H)$. The goal of the project is finding the proper form of the probability distribution P, the proper history vector, H, and the proper feature vector, x, and developing the infrastructure (e.g. efficient estimation and search techniques) so that consistency modeling can be effectively used.

During the first year of this effort, SRI focused on developing the appropriate base technologies for consistency modeling. The two most important accomplishments during the year were the development of Genone HMM technology, our choice for P above, and Progressive Search technology for HMM systems which allows to develop and use complex HMM formulations in an efficient manner. Other accomplishments included:

- Developing schemes for reducing time required to train HMM systems (training from alignments).
- Developing a scheme that efficiently implements cross-word acoustic models using Progressive Search technology.
- Development of an agglomerative scheme to cluster HMM state output distributions.
- Implementation of a discrete-density local consistency model, where output probabilities are conditioned on the previous observation, and the development of the necessary schemes for dealing with the explosion in the number of parameters caused by this model.
- Development of an information-theoretic framework for estimating the effect of the history H in the conditional HMM output distribution $P(x/s,H)$ when H is constrained to be one of the previous frames x_{t-i}.
- Implementation of continuous local-consistency modeling. Verified that the information-theoretic framework above indeed predicts recognition accuracy improvements, and measured performance of continuous local consistency for several different frame lags.
- Development of Gaussian-tree technology, which permits efficient computation of Gaussians by evaluating only those that have a reasonable chance of resulting in a good probability density function.

2.1 Genone-Based HMM technology

SRI has developed a new type of hidden Markov model speech recognition technique called genonic mixtures, or genones. In this type of system, Gaussian mixture components are shared among groups of states. These groupings are automatically determined using agglomerative clustering techniques. This technique automatically balances the modeling resolution/robustness trade-off depending on the amount of training data. Using this and other

techniques, SRI has reduced its word error rate on ARPA's November 1992 baseline 5,000 word Wall Street Journal bigram evaluation set from 13.0% to 7.7%, a change of 41%[1].

Genone modeling is important for consistency modeling because we plan to base our consistency modeling systems on conditional Gaussian output distributions. Because of limited training data, these high dimensional distributions will require careful balance between high resolution and robustness through parameter smoothing. This technique should permit us to achieve the best performance possible given the training data available. A paper describing this technique has been included in Appendix 1.

2.2 Progressive-Search Technology

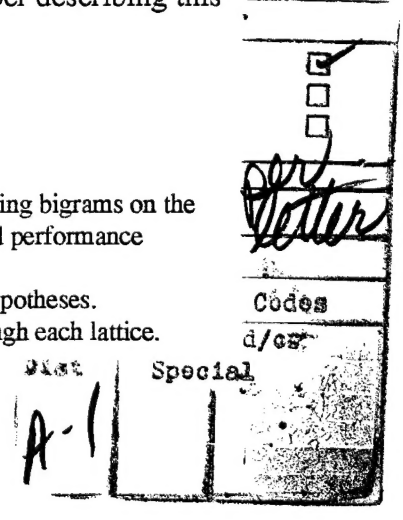
A technique called Progressive Search has been developed that allows speech recognition experiments for some of our computationally burdensome algorithms to be run over several hundred sentences in a few hours instead of a day or more. Progressive Search is a multiple-pass technique, with each pass using a progressively more accurate (and costly) algorithm. Each pass outputs a grammar (or word lattice) used to constrain the next pass's search space (instead of a less efficient N-best sentence list). It allows evaluation of computationally demanding algorithms (N-grams, more complex HMMs). It also facilitates developing real-time high-accuracy large-vocabulary recognition.

A Progressive Search technique has been applied to a standard cross-word tied-mixture 5K bigram HMM recognizer for ARPA's WSJ dictation task. It improved recognition development time by an order of magnitude (from 46 x Real Time to 5.6 x Real Time) when precomputed first-pass lattices were stored.

Another important application of the Progressive Search technique has been for trigram language models. In this case, the grammar output by an initial bigram-based recognizer was converted into a trigram grammar by replicating those states in the grammar where trigram word transition probabilities existed. This approach to trigram language modeling increased decoding time only slightly from that of bigram (15% increase), with only a minimal increase to the grammar size (since most of the trigrams were not represented). This approach is much more powerful than using an N-best approach to implementing trigram language models since more of the correct words exist in the lattice than the top N sentences. For instance, in a recent experiment using bigram language models for a 5,000-word WSJ speech recognition system, a system that achieved approximately a 10% word-error rate[2] on our development set achieved approximately 5% N-best error rate[3] for N = 1000, whereas the relatively compact grammar generated by this system had a 1% lattice error rate.[4] This reduced error rate gives the language model the opportunity to repair errors that the N-best system could not overcome. A paper describing this technique has been included in Appendix 2.

---

1. An error rate reduction of 25% was due solely to genone technology.
2. 10% bigram error on our development set is roughly equivalent to a 7% word error using bigrams on the official November 1992 evaluation set, approximately the same as the best bigram-based performance reported at the January 1993 ARPA meeting.
3. The N-best word error rate is defined as average error of the best of the N sentence hypotheses.
4. The lattice error rate is the average of the error rate associated with the best path through each lattice.



3. LISTS OF PUBLICATIONS, PRESENTATIONS AND REPORTS

3.1 Refereed papers submitted but not yet published:

Digalakis, V. and H. Murveit, "An Algorithm for Optimizing the Degree of Mixture-Tying in a Large-Vocabulary HMM-Based Speech Recognizer," submitted to the IEEE International Conference on Acoustics, Speech and Signal Processing, 1994.

L. Neumeyer, V. Digalakis, M. Weintraub, "Training Issues and Channel Equalization Techniques for the Construction of Telephone Acoustic Models Using a High-Quality Speech Corpus," to appear in IEEE Trans. Speech and Audio Processing, Special Issue, Spring 1994.

3.2 Refereed papers published[5]:

Murveit, H., J. Butzberger, V. Digalakis and M. Weintraub, "Large-Vocabulary Dictation Using SRI's DECIPHER(TM) Speech Recognition System: Progressive-Search Techniques," Proceedings ICASSP-93.

Murveit, H., J. Butzberger, V. Digalakis and M. Weintraub, "Progressive Search for Large Vocabulary Speech Recognition," Proceedings of the ARPA Human Language Technology Workshop, March 1993.

3.3 Invited presentations:

H. Murveit, "Progressive Search Techniques," ARPA Spoken Language Systems Technology Workshop, January 1993, Massachusetts Institute of Technology, Cambridge, Mass.

M. Weintraub, "SRI's Stress-Test Benchmark," ARPA Spoken Language Systems Technology Workshop, January 1993, Massachusetts Institute of Technology, Cambridge, Mass.

Demonstration of a 20,000-word continuous speech recognition in ARPA's Wall Street Journal domain, ARPA Spoken Language Systems Technology Workshop, January 1993, Massachusetts Institute of Technology, Cambridge, Mass.

M. Cohen, V. Digalakis, H. Murveit, P. Price, M. Weintraub, "Speech Recognition: an Overview, Examples and Demonstration," presented at Information Systems Laboratory, Stanford University, February 1993.

H. Murveit, Organized and gave overview and summary talk for the demonstration session of the ARPA Human Language Technology Workshop, March 22, 1993, Plainsboro, New Jersey.

M. Weintraub, "Progressive Search for Large Vocabulary Speech Recognition," ARPA Human Language Technology Workshop, March 22, 1993, Plainsboro, New Jersey.

Murveit, H., J. Butzberger, V. Digalakis and M. Weintraub, "Large-Vocabulary Dictation Using SRI's DECIPHER™ Speech Recognition System: Progressive-Search Techniques," ICASSP-93, April 1993.

V. Digalakis, "Search and Modeling Issues in Large-Vocabulary Speech Recognition" presented at Xerox PARC, August 1993.

---

5. Actually, these papers had extended abstracts that were refereed, the papers themselves were not refereed.

4. TRANSITIONS AND DoD INTERACTIONS

We were active participants in the 6-week Robust Speech Processing workshop sponsored by NSA at Rutgers in July-August 1993. Our two researchers there, Leo Neumeyer and Vassilios Digalakis, focussed on training issues and channel equalization techniques for acoustic modeling of telephone speech. Their work at the workshop will be included in a special issue of the IEEE Transaction on Speech and Audio Processing scheduled for publication in spring 1994.

SRI's Decipher[6] speech recognition technology is being transitioned to Boston University for joint research funded by NSF and ARPA, and we are currently arranging to modify our Decipher technology on SRI internal funds so that ARPA-sponsored research at CAIP in collaboration with Sarnoff Laboratories can take advantage of this technology for research on robust front-end signal processing. In addition, we are discussing with Nancy Chinchor at SAIC, the possibilty of using Decipher's technology in work to be conducted for ARPA and for NASA.

Several applications based on Decipher technology were demonstrated at Spoken Language Technology Applications Day last April. This event was attended by over 300 people, about equally divided among government and commercial representatives. Our participation in this event was sponsored by internal funds.

To further the transfer of Decipher technology, SRI has invested significant internal resources toward the development of robust, portable speech recognition software and tools for its use. Several commercial clients are using the resultant technology in their own research or in field trials.

---

6. Decipher is a trademark of SRI International.

5. SOFTWARE AND HARDWARE PROTOTYPES

The algorithms and software that is developed in this project will be incorporated into the Decipher speech recognition system. We are attempting to commercialize speech recognition based on Decipher and based on tools and other extensions to it that were funded by SRI International's IR&D support. SRI currently has several commercial clients that are in the process of evaluating speech recognition products based on Decipher.

Hy Murveit, SRI International
(415) 859-5447, hy@speech.sri.com,
High Performance Speech Recognition Using Consistency Modeling,
N00014-92-C-0154, 1 Oct 92 - 30 Sep 93

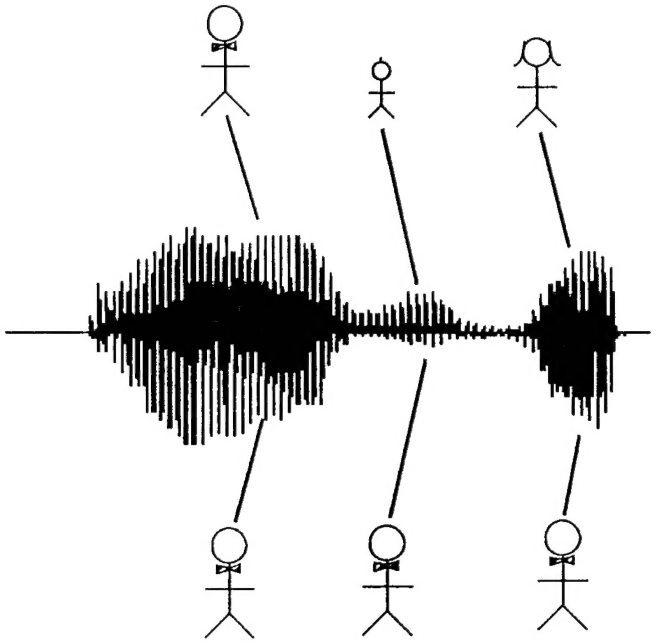
6. PHOTOGRAPHS, VUGRAPHS, AND VIDEOTAPES

See next page.

High-Performance Speech Recognition Using Consistency Modeling

SRI International

TRADITIONAL ALGORITHMS CAN USE INCONSISTENT MODELS



CONSISTENCY MODELING ALGORITHMS LEAD TO COHESIVE THEORIES

IMPACT

- Conventional simplifying independence assumptions in current speech recognition systems are clearly violated by speaker/microphone/noise variability. This makes speech recognition technology fragile.
- Consistency modeling is a breakthrough in technology and performance by providing a workable conceptual framework for handling these key problems.
- Resulting high-performance microphone-independent, speaker-adaptive, noise robust speech recognition systems will have a broad impact upon the application of speech understanding technology.

NEW IDEAS

- Remove local independence assumptions by computing the conditional density of the current feature given an appropriate representation of the previous feature vectors.
- Local consistency modeling computes the conditional probability of acoustic observations given the most salient features of recent acoustic observations.
- Global consistency modeling uses linguistic attributes estimated from the current or previous sentences to sharpen the output distributions.

SCHEDULE

- Evaluate Genone-HMM framework and limited global consistency modeling in ARPA's WSJ1 multi-site evaluation (completed 10/93).
- Continue to improve basic acoustic calibration techniques including improved probability distribution formulations and improved feature vectors (completed 12/93)
- Evaluate full global consistency techniques in a Genone-HMM framework using a variety of different history vectors (3/93)
- Extend search techniques for high speed global consistency modeling (6/94)
- Extend techniques to more difficult consistency modeling problems and evaluate performance (4/95)

APPENDICES

1) Digalakis, V., and H. Murveit, "An Algorithm for Optimizing the Degree of Mixture-Tying in a Large-Vocabulary HMM-Based Speech Recognizer," submitted to the *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994.*

2) Murveit, H., J. Butzberger, V. Digalakis and M. Weintraub, "Large-Vocabulary Dictation Using SRI's DECIPHER(TM) Speech Recognition System: Progressive-Search Techniques," *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993.*